BAB I
PENDAHULUAN

1.1 Latar belakang

Media sosial, khususnya Instagram, telah menjadi salah satu sarana utama masyarakat Indonesia dalam menyampaikan pendapat, kritik, maupun dukungan terhadap berbagai isu terkini karena banyaknya pengguna yang dapat mengungkapkan opini secara terbuka melalui fitur komentar (Pranata et al., 2022). Salah satu isu yang banyak menjadi perhatian masyarakat adalah mengenai gaji guru honorer. Isu tersebut muncul bersamaan dengan pemberitaan mengenai kenaikan tunjangan anggota DPR sehingga memunculkan berbagai tanggapan masyarakat terhadap perbedaan kondisi kesejahteraan di berbagai sektor. Di tengah pemberitaan tersebut, masih terdapat guru honorer yang menerima upah relatif rendah dan belum memperoleh kesejahteraan yang memadai. Kondisi tersebut mendorong masyarakat untuk menyampaikan berbagai pendapat melalui komentar di media sosial Instagram.

Banyaknya komentar yang muncul menyebabkan proses analisis secara manual menjadi kurang efektif karena membutuhkan waktu yang lama serta rentan terhadap subjektivitas. Selain itu, komentar-komentar tersebut mengandung emosi, ekspresi, dan penilaian pengguna terhadap suatu peristiwa sehingga diperlukan suatu pendekatan komputasi yang mampu mengidentifikasi kecenderungan opini masyarakat secara otomatis. Salah satu pendekatan yang banyak digunakan adalah analisis sentimen, yaitu proses mengelompokkan opini ke dalam kategori positif, negatif, maupun netral berdasarkan isi teks (Semary et al., 2024).

Dalam analisis sentimen, data yang digunakan masih berupa teks sehingga tidak dapat diproses secara langsung oleh algoritma *machine learning*. Oleh karena itu, diperlukan tahap representasi teks untuk mengubah data bahasa alami menjadi bentuk numerik agar dapat dipahami oleh algoritma komputer. Tahap representasi teks menjadi sangat penting karena memengaruhi kemampuan algoritma dalam mengenali pola bahasa pada data. Representasi teks yang kurang tepat dapat menyebabkan informasi yang terkandung dalam suatu dokumen tidak dapat direpresentasikan secara optimal sehingga memengaruhi performa klasifikasi (Mutinda et al., 2023).

Salah satu metode representasi teks yang banyak digunakan adalah *Word2Vec*. *Word2Vec* mengubah kata menjadi vektor berukuran kecil berdasarkan hubungan kemunculan kata dalam suatu konteks sehingga kata-kata yang ditemukan pada konteks yang serupa akan memiliki representasi yang saling berdekatan (Worth, 2023). Metode ini memberikan kemampuan bagi model untuk memahami hubungan kontekstual setiap kata yang direpresentasikan ke dalam ruang vektor berkelanjutan (Alami et al., 2019). Namun demikian, *Word2Vec* memiliki keterbatasan karena kualitas representasi yang dihasilkan sangat bergantung pada kualitas dan jumlah korpus pelatihan. Apabila data yang digunakan kurang bervariasi, hubungan makna yang dipelajari model menjadi kurang optimal (Nur et al., 2025).

Selain *Word2Vec*, metode representasi teks yang banyak digunakan adalah *N-Gram* . Metode ini membentuk fitur berdasarkan urutan kata dalam rentang tertentu, seperti *Unigram*, *Bigram*, dan *Trigram* (Muria et al., 2023). *N-Gram* relatif mudah diterapkan dan mampu menangkap pola kemunculan kata maupun frasa yang sering muncul dalam suatu dokumen (Supriadi et al., 2020). Akan tetapi, metode ini memiliki kelemahan berupa meningkatnya jumlah fitur secara signifikan ketika ukuran dataset maupun nilai *n* semakin besar (Sujaini & Putra, 2024). Selain itu, kata-kata yang memiliki makna serupa tetapi muncul dalam urutan yang berbeda tetap dianggap sebagai fitur yang berbeda karena metode ini belum mampu memahami hubungan semantik antarkata (Wardani & Evangelista, 2024).

Berdasarkan karakteristik tersebut, *Word2Vec* dan *N-Gram* memiliki kelebihan serta keterbatasan masing-masing sehingga diperlukan analisis lebih lanjut untuk mengetahui metode representasi teks yang mampu menghasilkan performa klasifikasi yang lebih baik. Setelah proses representasi teks dilakukan, tahap selanjutnya adalah klasifikasi sentimen menggunakan algoritma *machine learning*. Pada penelitian ini digunakan algoritma *Support Vector Machine* (SVM) karena mampu melakukan analisis pada data berdimensi tinggi dan mencari *hyperplane* optimal sebagai pemisah antar kelas sehingga banyak digunakan dalam klasifikasi teks (Cinta Amaria & Hidayati, 2025). Selain itu, seperti dijelaskan oleh Andriansyah dan Astuti (2025), SVM tidak dapat memproses data teks mentah secara langsung sehingga memerlukan tahap representasi fitur sebelum proses

klasifikasi dilakukan. Di samping metode representasi teks, performa SVM juga dipengaruhi oleh fungsi kernel yang digunakan. Kernel berfungsi mentransformasikan data ke ruang fitur berdimensi lebih tinggi sehingga data yang tidak dapat dipisahkan secara linear dapat diklasifikasikan dengan lebih baik. Oleh karena itu, penelitian ini menggunakan empat fungsi kernel, yaitu *Linear, Polynomial, Radial Basis Function* (RBF), dan *Sigmoid*, untuk mengetahui kernel yang memberikan performa terbaik pada masing-masing metode representasi teks.

Berdasarkan uraian tersebut, penelitian ini menggunakan dataset berupa komentar masyarakat pada media sosial Instagram yang membahas isu gaji guru honorer untuk membandingkan metode representasi teks *Word2Vec* dan *N-Gram* terhadap kinerja algoritma *Support Vector Machine* menggunakan kernel *Linear, Polynomial, Radial Basis Function* (RBF), dan *Sigmoid* dalam klasifikasi sentimen. Penelitian ini diharapkan dapat memberikan informasi mengenai metode representasi teks dan fungsi kernel yang menghasilkan performa klasifikasi terbaik pada klasifikasi sentimen komentar Instagram terkait gaji guru honorer.

1.2 Rumusan Masalah

Berdasarkan uraian pada latar belakang, maka rumusan masalah pada penelitian ini adalah bagaimana kinerja algoritma *Support Vector Machine* dalam mengklasifikasikan sentimen positif, negatif atau netral pada komentar Instagram dengan menggunakan *Word2Vec* dan *N-Gram* sebagai representasi fitur?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dipaparkan, maka tujuan dari penelitian ini adalah mengetahui kinerja algoritma *Support Vector Machine* dalam mengklasifikasikan sentimen positif, negatif atau netral pada komentar Instagram terkait isu gaji guru honorer dengan menggunakan *Word2Vec* dan *N-Gram* sebagai metode representasi teks, yang diukur berdasarkan nilai akurasi, presisi, recall, dan f1-score

1.4 Manfaat Penelitian

Berdasarkan tujuan penelitian, hasil dari penelitian ini diharapkan mampu memberikan manfaat, baik dari sisi teoritis maupun praktis yaitu sebagai berikut:

1) Manfaat Teoritis :

Diharapkan bahwa melalui penelitian ini mampu membantu kemajuan ilmu pengetahuan dalam bidang ekstraksi informasi dari teks dan *Natural Language Processing* (NLP), terutama penerapan perbandingan metode representasi teks *Word2Vec* dan *N-Gram* dalam meningkatkan hasil algoritma *Support Vector Machine* (SVM) dalam proses klasifikasi sentimen berbahasa Indonesia.

2) Manfaat Praktis
   a. Bagi peneliti, dapat digunakan sebagai bahan pertimbangan untuk memahami bagaimana menerapkan kedua teknik *Word2Vec* dan *N-Gram* secara efektif sehingga meningkatkan tingkat akurasi dalam klasifikasi sentimen.
   b. Bagi pengembang sistem, penelitian ini diharapkan dapat menjadi dasar merancang sistem analisis berbasis kecerdasan buatan yang lebih tepat dan akurat dalam konteks bahasa informal di media sosial.

1.5 Batasan Penelitian

Untuk menjaga penelitian ini agar tetap terarah dari tujuan yang telah ditetapkan, pembatasan masalah diperlukan sebagai ruang lingkup data yang akan digunakan. Berikut adalah penjelasannya:

1) Penelitian ini akan berfokus pada komentar pengguna Instagram yang membahas isu guru honorer
2) Data diambil dari Instagram dari tanggal 14 Juli 2025 – 1 November 2025
3) Data awal yang diambil 31.880 data komentar Instagram
4) Data yang digunakan setelah filtering manual 17.263 data komentar Instagram
5) Data yang akan digunakan dalam sebuah penelitian ini adalah komentar - komentar publik dari dua postingan unggahan Instagram tertentu.
6) Evaluasi dari model dilakukan pada parameter seperti akurasi, presisi, recall, dan *F1-Score* digunakan untuk menilai hasil klasifikasi.
7) Metode pengambilan data menggunakan teknik *web scrapping* berbasis *library Instagrapi* dengan Python
8) Penanganan ketidakseimbangan data dilakukan menggunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE).

9) Algoritma klasifikasi yang digunakan adalah *Support Vector Machine* (SVM) dengan kernel *Linear, Polynomial, Radial Basis Function* (RBF), dan *Sigmoid*.